Multi-Granular Multimodal Clue Fusion for Meme Understanding

Li Zheng[1], Hao Fei[2], Ting Dai[1], Zuquan Peng[1], Fei Li[1,3*], Huisheng Ma[4], Chong Teng[1], Donghong Ji[1]

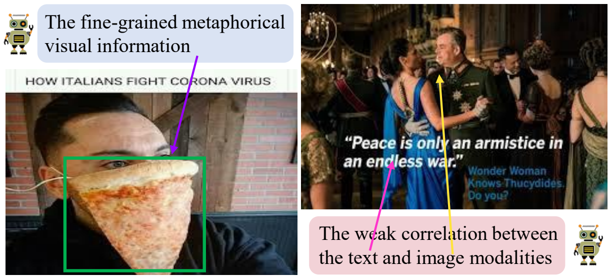
[1]Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education,
School of Cyber Science and Engineering, Wuhan University, Wuhan, China
[2]National University of Singapore, Singapore, Singapore
[3]Laboratory for Advanced Computing and Intelligence Engineering, Wuxi, China
[4]North China Institute of Computing Technology, beijing, China
{zhengli,daiting_cs,pzq_cse,lifei_csnlp,tengchong,dhji}@whu.edu.cn
haofei37@nus.edu.sg, mhs@bupt.cn

Abstract

With the continuous emergence of various social media platforms frequently used in daily life, the multimodal meme understanding (MMU) task has been garnering increasing attention. MMU aims to explore and comprehend the meanings of memes from various perspectives by performing tasks such as metaphor recognition, sentiment analysis, intention detection, and offensiveness detection. Despite making progress, limitations persist due to the loss of fine-grained metaphorical visual clue and the neglect of multimodal text-image weak correlation. To overcome these limitations, we propose a multi-granular multimodal clue fusion model (MGMCF) to advance MMU. Firstly, we design an object-level semantic mining module to extract object-level image feature clues, achieving fine-grained feature clue extraction and enhancing the model's ability to capture metaphorical details and semantics. Secondly, we propose a brand-new global-local cross-modal interaction model to address the weak correlation between text and images. This model facilitates effective interaction between global multimodal contextual clues and local unimodal feature clues, strengthening their representations through a bidirectional cross-modal attention mechanism. Finally, we devise a dual-semantic guided training strategy to enhance the model's understanding and alignment of multimodal representations in the semantic space. Experiments conducted on the widely-used MET-MEME bilingual dataset demonstrate significant improvements over state-of-the-art baselines. Specifically, there is an 8.14% increase in precision for offensiveness detection task, and respective accuracy enhancements of 3.53%, 3.89%, and 3.52% for metaphor recognition, sentiment analysis, and intention detection tasks. These results, underpinned by in-depth analyses, underscore the effectiveness and potential of our approach for advancing MMU.

Introduction

Memes, as a popular form of online communication, express viewpoints, sentiments, and intentions in a concise and humorous manner. With the development of social networks, Multimodal Meme Understanding (MMU) (Wang et al. 2024; Xu et al. 2022), as an emerging research area in Natural Language Processing (NLP), plays a crucial role in many downstream applications, such as question answering (Zheng et al. 2024b) and sentiment analysis (Zheng et al. 2023a,b). The definition of the MMU task involves predicting understanding from four dimensions: metaphor, sentiment, intention, and offensiveness. However, memes are nuanced, and accurately grasping the underlying meaning embedded within the combination of text and images poses a crucial challenge.


The fine-grained metaphorical visual information
HOW ITALIANS FIGHT CORONA VIRUS
"Peace is only an armistice in an endless war."
Wonder Woman Knows Thucydides. Do you?
The weak correlation between the text and image modalities

(a) How Italians fight corona virus?

(b) Peace is only an armistice in an endless war.

Figure 1: Examples of Metaphorical Memes.

Several studies have made commendable efforts in MMU. Kiela et al. (2020); Kirk et al. (2021) introduced multimodal hate meme datasets specifically designed for hate detection. However, these studies overlooked the crucial aspect of metaphorical features in memes. Therefore, Xu et al. (2022) considered the richer metaphorical features in memes and constructed a baseline model and a bilingual dataset called MET-MEME for this purpose. Furthermore, Wang et al. (2024) proposed a metaphor-aware multimodal multitask framework on this dataset to capture the interactions between text and images. Despite achieving notable success, current researches in this field face two significant limitations: 1) **the loss of fine-grained metaphorical visual clues** and 2) **the neglect of multimodal text-image weak correlation**. These limitations hinder its further flourishing and widespread adoption.

On the one hand, existing works (Qu et al. 2023; Ji, Ren, and Naseem 2023) exhibit a lack of emphasis on images and simply encode broad visual representations at the image-level, ignoring metaphorical clues at the fine-grained object-level of images. This neglect leads to a critical absence of key visual metaphorical details, resulting in semantic ambiguity and omissions, ultimately failing to comprehensively capture

[*]Corresponding author.

the complexity and diversity of memes. As shown in Figure 1 (a), encoding visual features solely at the image-level falls short in capturing the crucial metaphorical clues of a pizza being used as a mask. Detecting the presence of metaphors and accurately predicting the conveyed sentiments, intentions, and potential offensiveness becomes exceedingly challenging in such cases.

On the other hand, existing methods (Xu et al. 2022; Wang et al. 2024) primarily focus on directly integrating textual and visual information to comprehend memes, overlooking the issue of weak correlations between modalities and disregarding the intrinsic crucial metaphorical clues within each modality. This oversight lead to the loss of crucial details and clues, thereby limiting a comprehensive understanding of memes. For instance, in the illustrated example in Figure 1 (b), there is a weak correlation between the image and text, where the image conveys peace while the text reveals hatred towards war. Merely concatenating and fusing the image and text information directly could lead to a misinterpretation of the meme as peaceful.

In this paper, motivated by the aforementioned observations, we propose a ***M**ulti-**G**ranular **M**ultimodal **C**lue **F**usion model (MGMCF)* to improve multimodal meme understanding. **First**, we design an object-level semantic mining module to extract fine-grained object-level feature clues from images. We then integrate these object-level feature clues with the overall image-level feature clues to obtain a multi-granular representation. This enables our model to better capture the metaphorical details and semantics of images, offering a more comprehensive visual understanding. **Second**, considering the weak correlation between text and images, we not only focus on the interactions between different modalities but also emphasize the crucial metaphorical clues within each modality, integrating multi-granular clues to enhance the ultimate understanding of multimodal memes. Therefore, we propose a novel global-local cross-modal interaction model to enable effective interaction between the global multimodal contextual clues and local unimodal feature clues. Specifically, the global multimodal context enhances the local unimodal features through a symmetric cross-modal attention mechanism. This interaction process is bidirectional, allowing the global context to extract useful clues from the local unimodal features and update itself. Through multi-level stacking, the global multimodal context and local unimodal features mutually enhance each other and gradually improve. **Moreover**, to enhance semantic alignment, we devise a dual-semantic guided training strategy. By bringing related image-text pairs closer in the forward direction and pushing unrelated pairs apart in the reverse direction, we aim to foster a more robust understanding of complex multimodal semantic clues.

To verify the effectiveness of our model, we conduct experiments on the benchmark MET-MEME bilingual dataset (Xu et al. 2022), which contains both English and Chinese memes. The results demonstrate that our model significantly outperforms all state-of-the-art (SoTA) baselines across all evaluation metrics in the four tasks. On the English meme dataset, the precision in the offensiveness detection task increased by 8.14%, and the accuracy in metaphor recognition, sentiment analysis, and intention detection tasks improved by 3.53%, 3.89%, and 3.52%, respectively. Additionally, extensive experiments validate the effectiveness of our fine-grained visual information enhancement and global-local interactions.

Our main contributions are summarized as follows:

- We analyze and summarize two intrinsic challenges in the MMU task and propose a multi-granular multimodal clue fusion model, the first to consider fine-grained visual clues and integrate unimodal feature clues to enhance MMU.
- We design an object-level semantic mining module and a global-local cross-modal interaction model to facilitate effective interaction between global multimodal clues and local unimodal clues, achieving multi-granular understanding of meme metaphorical clues and semantics.
- Our extensive experimental results on MET-MEME dataset demonstrate that our scheme achieves state-of-the-art performance on the MMU task.

Related Work

Multimodal Meme Understanding

Recently, multimodal meme understanding (Lin et al. 2024; Hee, Chong, and Lee 2023; Qu et al. 2023; Fang et al. 2024b, 2023, 2024c, 2025; Zheng et al. 2024a) has attracted increasing attention. Unlike general multimodal learning tasks (Ji et al. 2021, 2022; Li et al. 2022b,a; Wu et al. 2023; Li et al. 2023a; Fei et al. 2024a; Luo et al. 2024), meme understanding relies more heavily on contextual and metaphorical information. Existing research has mainly focused on hateful memes. Yang et al. (2023) proposed a scalable invariant and specific modality representation learning framework based on graph neural networks for harmful meme detection. Ji, Ren, and Naseem (2023) introduced a prompt-based method to identify harmful memes. Beyond just focusing on hateful meme detection, Xu et al. (2022) introduced a fine-grained multimodal meme understanding dataset, which includes tasks such as sentiment analysis, intention detection, offensiveness detection, and metaphor recognition, to analyze memes at a finer granularity. Wang et al. (2024) created cross-modal and intra-modal attention mechanisms on this dataset to capture the interactions between text and images for multimodal meme understanding. However, existing works overlook the object-level fine-grained clues in images, potentially leading to the loss of crucial metaphorical visual details. Moreover, these methods do not address the issue of cross-modal weak correlations, neglecting the essential clues within unimodal, which result in semantic ambiguity and confusion, failing to fully capture the complexity and diversity of memes.

Metaphorical Information Processing

In the field of NLP, there has been growing interest in developing models for metaphor detection (He et al. 2024; Fang et al. 2024a; Elzohbi and Zhao 2024; Zhang and Liu 2023). Understanding the essence of memes relies critically on identifying the metaphorical information embedded within them. Existing researches (Qiao, Zhang, and Ma 2024; Elzohbi and Zhao 2024; Badathala et al. 2023) have primarily focused on unimodal metaphor detection. Zhang and Liu (2023) proposed a novel multi-task learning framework based on a metaphor

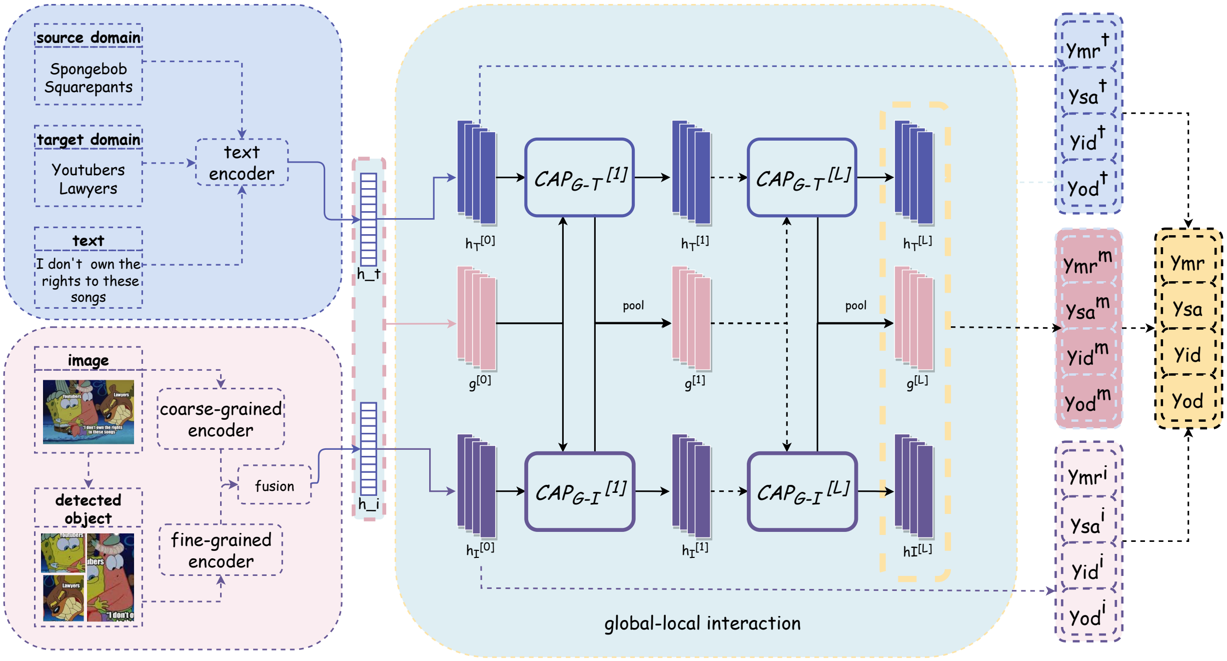


Figure 2: The overall architecture of our model. “mr” means metaphor recognition, “sa” means sentiment analysis, “id” means intention detection, “od” means offensiveness detection.

recognition program, a set of linguistic rules. Li et al. (2023b) performed metaphor detection by explicitly modeling the basic meanings of concepts. Tian et al. (2023) designed a domain contrastive learning strategy to capture the semantic inconsistencies. While these unimodal metaphor detection methods have achieved promising results, there has been relatively less exploration in the area of multimodal metaphor detection. Alnajjar, Hämäläinen, and Zhang (2022) introduced a multimodal metaphor annotated corpus and designed a video-text content-based method for metaphor detection. He et al. (2024) developed a multi-interactive cross-modal residual network for multimodal metaphor recognition.

Methodology

Task Definition

This paper addresses the task of multimodal meme understanding, encompassing metaphor recognition, sentiment analysis, intention detection, and offensiveness detection. Specifically, given an example consisting of an image I, its corresponding text T, a source domain T_s, and a target domain T_a, the objective of multimodal meme analysis is to predict the categories of metaphor y_{mr}, sentiment y_{sa}, intention y_{id}, and offensiveness y_{od}. As shown in Figure 2, the source domain serves as the basis for the metaphor, while the target domain embodies the concept or idea metaphorically conveyed, typically in textual form.

Feature Extraction

Text Encoder. In accordance with the approach described in Wang et al. (2024), we utilize Multilingual BERT (Wang et al. 2019) to extract textual features from the corresponding text x^t, source domain x^s, and target domain x^a. The encoding process can be formulated briefly as:

$$\begin{aligned}\{\boldsymbol{h}_1^t, ..., \boldsymbol{h}_n^t\} &= M\text{-}BERT(\{\boldsymbol{x}_1^t, ..., \boldsymbol{x}_n^t\})\\ \{\boldsymbol{h}_1^s, ..., \boldsymbol{h}_p^s\} &= M\text{-}BERT(\{\boldsymbol{x}_1^s, ..., \boldsymbol{x}_p^s\})\\ \{\boldsymbol{h}_1^a, ..., \boldsymbol{h}_q^a\} &= M\text{-}BERT(\{\boldsymbol{x}_1^a, ..., \boldsymbol{x}_q^a\})\end{aligned} \tag{1}$$

where n, p, and q represent the word counts of the corresponding text, source domain, and target domain, respectively.

Image Encoder.

Multimodal meme images contain rich metaphorical details that are crucial clues for understanding memes. Therefore, when extracting visual features from memes, we cannot solely focus on image-level visual semantic clues as with other multimodal tasks. It is imperative to capture object-level fine-grained clue features that encompass these metaphorical details. To achieve this goal, we devise a visual information enhancement strategy for extracting feature clues of different granularities.

For image I, following (Wang et al. 2024), we first employ a pretrained convolutional neural network classifier, VGG16 (Simonyan and Zisserman 2014), to extract image-level features $\boldsymbol{h}^c = VGG16(I)$. Then, we transform the input image I into a series of embedded blocks to capture fine-grained image features. By integrating object detection, attribute recognition, and positional information, we enrich the representation of image features and enhance enhance image metaphor comprehension. Specifically, we design an object-level semantic mining module (Anderson et al. 2018) to identify and localize objects in an image. For each visual region I_i represented by a bounding box, we resize the region to a standard size of 224× 224 pixels. Following Xu, Zeng, and Mao (2020), we reshape the resized region I_i into a sequence $I_i = \{r_1, ..., r_m\}$, where each region is represented by a block. This reshaping divides the region into a grid of blocks, with m being the total number of blocks. Next, we flatten each block r_j and project it into a d^I-dimensional vector. This projection is performed using a trainable linear projection matrix E, and the resulting embedded representation of block r_j is denoted as $z_j = r_j E$. To incorporate contextual information and retain positional information, we prepend a [class] token embedding at the beginning of the patch sequence. Position embeddings are also appended to the patch embeddings, indicating their relative positions within the sequence. The input representation of each visual region I_i is expressed as:

$$\boldsymbol{Z}_i = [\boldsymbol{z}_{[class]}; \boldsymbol{z}_1, ..., \boldsymbol{z}_m] + \boldsymbol{E}_{pos} \tag{2}$$

where Z_i represents the input matrix of image patches, and E_{pos} denotes the position embedding matrix. Subsequently, we feed the input matrix Z_i into the VGG16 encoder to obtain the visual region I_i representation $h_i^v = VGG16(Z_i)$. Finally, the representation of the image I is defined as:

$$\boldsymbol{h}_I = \{\boldsymbol{h}^c, \boldsymbol{h}_1^v, ..., \boldsymbol{h}_m^v\} \tag{3}$$

Modal Fusion

The text and image of multimodal meme have the problem of weak correlation, and directly fusing the text and image may result in incorrect meme understanding. A good fusion solution should extract and integrate sufficient information from multimodal sequences while preserving the independence of each modality. Therefore, we propose a novel global-local cross-modal interaction model that not only considers interactions between modalities but also emphasizes the importance of each modality itself to enhance multimodal fusion at multiple granularities. Specifically, we devise an efficient mechanism called Cross-modal Attention Promotion (CAP) that leverages symmetric cross-modal attention to explore the inherent correlations between the two input feature sequences, promoting the exchange of beneficial information across the sequences. CAP utilizes self-attention to model the temporal dependencies within each feature sequence, enabling the integration of more information. The mechanism takes sequences $\boldsymbol{h}^T$ and $\boldsymbol{h}^I$ as inputs and generates their mutually reinforcing information $\boldsymbol{h}_{T\to I}$ and $\boldsymbol{h}_{I\to T}$. The computation of $CAP_{T\leftrightarrow I}(\boldsymbol{h}^T, \boldsymbol{h}^I)$ is as follows:

$$\begin{aligned} \boldsymbol{h}'_{T\to I} &= MCA(LN(\boldsymbol{h}_T), LN(\boldsymbol{h}_I)) + \boldsymbol{h}_T \\ \boldsymbol{h}''_{T\to I} &= MSA(LN(\boldsymbol{h}'_{T\to I})) + \boldsymbol{h}'_{T\to I} \\ \boldsymbol{h}_{T\to I} &= FN(LN(\boldsymbol{h}''_{T\to I})) + \boldsymbol{h}''_{T\to I} \end{aligned} \tag{4}$$

where LN denotes layer normalization and FN is the feedforward neural network. $MSA(\cdot)$ refers to the output of the multi-head self-attention mechanism computation. $MCA(\cdot,\cdot)$represents the result of the multiple cross-attention mechanism calculation. Similarly, we can obtain $CAP_{I\leftrightarrow T}(\boldsymbol{h}_I, \boldsymbol{h}_T)$.

The traditional cross-attention interaction requires two updates during the modal interaction process to achieve modal enhancement, which is inefficient and introduces redundant features into the sequence. Based on the historical experience from large-scale pretraining, it has been observed that a single token can represent the entire sequence, further improving the efficiency of modal interaction. Motivated by this observation, we propose a global-local cross-modal interaction model with linear computational cost to enhance efficiency. The discourse-level representation of each modality replaces the standard information and interacts with local unimodal features within a global multimodal context. This means that the representation of each modality not only relies on local features but also takes into account the influence of global context. This global-local interaction model reduces the introduction of redundant features and improves modal interaction effectiveness while maintaining efficiency.

We establish the global multimodal context information denoted as $\boldsymbol{g}^i = concat(\boldsymbol{h}_T^i, \boldsymbol{h}_I^i)$ by concatenating the representations of each modality at each layer of global-local interaction, where i represents the number of layers of global local interaction. By integrating the global context information and local modal information, and learning modal consistency and specificity, we ensure effective interaction and capture relevant information from both the global and local perspectives. The entire interaction process is as follows:

$$\begin{aligned} \boldsymbol{h}_T^{(i+1)}, \boldsymbol{g}_{T\to G}^{(i)} &= CAP_{T\leftrightarrow G}^{(i)}(\boldsymbol{h}_T^{(i)}, \boldsymbol{g}^{(i)}) \\ \boldsymbol{h}_I^{(i+1)}, \boldsymbol{g}_{I\to G}^{(i)} &= CAP_{I\leftrightarrow G}^{(i)}(\boldsymbol{h}_I^{(i)}, \boldsymbol{g}^{(i)}) \end{aligned} \tag{5}$$

By stacking multiple layers, the global multimodal context and local unimodal features can mutually reinforce and progressively refine each other. We hierarchically handle the entire learning process, with each layer capturing different features corresponding to the model's main stages. The model initially learns shallow interaction features, gradually progressing to acquire higher-order semantic features in later stages. This hierarchical learning method successfully integrates information from various modalities by ingeniously designed aggregation blocks, providing the model with a more comprehensive and enriched representation of multimodal features. Through the model's interactions, information from different modalities can be combined in a deeper and more effective manner, enabling the acquisition of more advanced feature representations in subsequent hierarchical learning. Subsequently, we aggregate the features from both unimodal and multimodal sources to facilitate subsequent task predictions.

$$y_m = softmax(\boldsymbol{W}_m MSA([\boldsymbol{h}_T^{(L)}, \boldsymbol{h}_I^{(L)}, \boldsymbol{g}^{(L)}]) + \boldsymbol{b}) \tag{6}$$

where y_m is the feature output distribution after multimodal fusion, W_m and b are trainable parameters.

Our approach focuses not only on the interactions between modalities but also on the individual feature representations of each modality. We separately use the unimodal features obtained from text and image encoders to predict subsequent tasks. This allows to capture the unique characteristics and information within each modality, thereby improving the accuracy and effectiveness of MMU.

$$\begin{aligned} \boldsymbol{y}_T &= softmax(\boldsymbol{W}_t \boldsymbol{h}_T + \boldsymbol{b}) \\ \boldsymbol{y}_I &= softmax(\boldsymbol{W}_i \boldsymbol{h}_I + \boldsymbol{b}) \end{aligned} \tag{7}$$

Given the y_M, y_T, and y_I, we obtain the final prediction y:

$$\boldsymbol{y} = \boldsymbol{y}_M + \boldsymbol{y}_T + \boldsymbol{y}_I \tag{8}$$

where y can be considered as a comprehensive feature set encompassing multi-granular features, including text, image, and image-text modalities.

Method	English			Chinese			English			Chinese		
	Acc	Pre	Rec	Acc	Pre	Rec	Acc	Pre	Rec	Acc	Pre	Rec
	Sentiment Analysis						Intention Detection					
VGG16	20.57	20.84	24.22	29.94	26.04	29.20	37.19	38.71	38.15	47.48	49.21	47.81
DenseNet-161	21.88	21.71	25.65	29.45	27.50	29.36	38.10	39.31	37.89	47.23	39.24	47.06
ResNet-50	21.74	18.63	21.35	29.36	27.50	29.28	39.19	37.12	40.10	47.15	39.23	47.06
Multi-BERT_EfficientNet	28.52	24.52	29.04	33.50	35.29	33.42	43.10	41.54	42.19	51.03	43.06	51.03
Multi-BERT_ViT	24.43	23.41	23.96	33.25	27.33	32.84	41.28	40.13	40.62	50.62	41.32	50.62
Multi-BERT_PiT	25.00	27.82	28.12	33.66	33.58	33.09	42.23	41.09	41.02	50.21	50.00	50.04
MET_add	24.65	24.52	25.26	32.50	32.62	33.50	40.32	40.39	41.28	52.93	52.68	54.01
MET_cat	27.68	28.41	29.82	33.42	34.33	33.91	38.56	39.19	39.84	51.58	51.48	52.85
M3F_add	30.47	33.45	30.34	39.95	41.80	39.87	44.40	41.89	44.32	55.25	54.57	55.00
M3F_cat	29.82	34.18	30.73	37.22	39.55	37.97	44.10	44.56	43.53	53.52	54.72	54.52
Ours	**34.36** (+3.89%)	**37.77** (+3.52%)	**34.38** (+3.65%)	**42.11** (+2.16%)	**47.59** (+5.79%)	**42.02** (+2.15%)	**47.92** (+3.52%)	**47.53** (+2.97%)	**47.06** (+2.74%)	**58.56** (+3.31%)	**57.99** (+3.27%)	**58.32** (+3.32%)
	Offensiveness Detection						Metaphor Recognition					
VGG16	67.10	63.42	72.53	70.07	64.11	72.07	78.39	79.73	79.95	67.00	67.24	67.82
DenseNet-161	69.66	62.07	69.98	71.43	70.82	75.43	80.08	80.23	80.47	67.16	67.91	67.99
ResNet-50	69.21	64.62	72.57	73.24	69.62	75.74	80.34	81.22	80.86	67.74	67.63	67.66
Multi-BERT_EfficientNet	73.78	67.98	74.56	78.15	72.11	79.98	82.46	84.39	83.11	74.19	71.26	74.28
Multi-BERT_ViT	71.22	62.96	72.66	76.92	67.31	78.74	81.90	82.01	82.46	73.28	72.55	73.70
Multi-BERT_PiT	72.79	66.69	74.26	77.17	70.16	79.05	82.07	83.05	82.98	75.10	73.15	74.28
MET_add	68.39	66.21	72.14	76.01	74.76	78.16	81.33	81.49	82.29	74.04	74.51	74.96
MET_cat	67.25	66.15	74.48	73.19	71.59	79.49	82.39	82.69	83.33	72.90	72.80	75.67
M3F_add	76.17	69.45	76.19	80.81	76.00	80.73	83.98	85.86	84.38	77.01	72.94	82.68
M3F_cat	74.09	69.59	76.15	80.07	76.20	80.62	83.20	85.97	85.81	76.18	73.02	80.00
Ours	**78.11** (+1.94%)	**77.73** (+8.14%)	**78.32** (+2.13%)	**82.15** (+1.34%)	**81.80** (+5.6%)	**82.02** (+1.29%)	**87.51** (+3.53%)	**88.58** (+2.61%)	**88.89** (+3.08%)	**78.39** (+1.38%)	**78.16** (+3.65%)	**83.92** (+1.24%)

Table 1: Experimental results on the MET-MEME dataset of four tasks.

Training

For each task, we train the model using the standard gradient descent algorithm to minimize the cross-entropy loss.

$$\min_{\theta} \mathcal{L}_* = -\sum_{i=1}^{N} y_*^i log \hat{y}_*^i + \lambda_* \left\|\theta_*\right\|^2 \tag{9}$$

where $*$ stands for the representation of different tasks. N is the training data size. y^i and $\hat{y}^i$ respectively represent the ground-truth and estimated label distribution of instance i. θ denotes all trainable parameters of the model, λ represents the coefficient of L2-regularization.

Dual-semantic Guided Loss. In addition to task-specific losses, we devise a dual-semantic guided loss to effectively leverage cross-modal information, enhancing the model's comprehension and alignment of multimodal representations in the semantic space. The context-aware multimodal representations ($\boldsymbol{h}T_i$ and $\boldsymbol{h}I_I$) contain contextually relevant information associated with specific memes. By bringing related image-text pairs closer in the forward direction and pushing unrelated pairs apart in the reverse direction, these representations are aligned in the same semantic space, effectively utilizing cross-modal information. Specifically, we contrast the multimodal representation of a specific meme sample (i.e., $\boldsymbol{h}_{T_i}$), with another multimodal representation ($\boldsymbol{h}_{I_i}$) from within the same batch of sampled memes. By comparing the similarities and differences between these representations, the model learns how to better differentiate and capture the semantic information among different meme samples.

$$\mathcal{L}_{dg} = -log \frac{exp(sim(\boldsymbol{h}_{T_i}, \boldsymbol{h}_{I_i})/\tau)}{\sum_{k=1[k \neq i]}^{2N} exp(sim(\boldsymbol{h}_{T_k}, \boldsymbol{h}_{I_k})/\tau))} \tag{10}$$

where sim is the cosine-similarity, N is the batch size, and τ is the temperature to scale the logits.

By minimizing task-specific losses for each individual task and incorporating a contrastive loss, our overall loss function is defined as follows:

$$\mathcal{L} = \mathcal{L}_{mr} + \mathcal{L}_{sa} + \mathcal{L}_{id} + \mathcal{L}_{od} + \mathcal{L}_{dg} \tag{11}$$

Experiments

Experimental Setting

Datasets. We assess the efficacy of our model on the widely used MET-MEME bilingual dataset (Xu et al. 2022), which consists of both English and Chinese memes. The English meme dataset is sourced from MEMOTION and Google search, comprising 4,000 text-image pairs, comprising 4,000 text-image pairs. The Chinese meme dataset consists of 6,045 text-image pairs, covering six different categories, including animals, scenery, animations, films, dolls, and humans.

Evaluation Metrics. In terms of evaluation metrics, we align with Wang et al. (2024) and use three metrics, namely accuracy (Acc), weighted precision (Pre), and recall (Rec), to assess the performance. All our scores are the average over 5 runnings with random seeds.

Baseline Systems

To validate the effectiveness of our model, we compare it against the following state-of-the-art baselines. (1) Unimodal models that solely utilize image modality information, such as VGG16 (Simonyan and Zisserman 2014), DenseNet-161 (Huang et al. 2017), and ResNet-50 (He et al. 2016). (2) Multimodal models that incorporate both text and image modalities. These include Multi-BERT-EfficientNet (Tan and Le 2019), Multi-BERT-ViT (Dosovitskiy et al. 2020), Multi-BERT-PiT (Heo et al. 2021), MET (Xu et al. 2022) employ a straightforward concatenation or element-wise addition to

Method	Sentiment Analysis			Intention Detection		
	Acc	Pre	Rec	Acc	Pre	Rec
Ours	34.36	37.77	34.38	47.92	47.53	47.06
w/o OM	32.47	36.16	32.39	46.27	46.17	45.89
w/o UP	32.85	36.53	32.76	46.73	46.59	46.11
w/o GL	31.76	35.49	31.83	45.82	45.41	45.47
w/o DG	33.59	36.94	33.52	47.05	46.94	46.63
	Offensiveness Detection			**Metaphor Recognition**		
	Acc	Pre	Rec	Acc	Pre	Rec
Ours	78.11	77.73	78.32	87.51	88.58	88.89
w/o OM	77.28	73.41	77.22	85.42	87.45	87.46
w/o UP	77.53	74.66	77.68	85.81	87.64	87.74
w/o GL	76.89	72.57	76.94	84.98	86.77	86.79
w/o DG	77.74	75.39	77.93	86.43	87.93	88.02

Table 2: Ablation results on the MET-MEME English dataset of four tasks. "OM" means object-level semantic mining, "UP" means unimodal prediction, "GL" means global-local interaction, "DG" means dual-semantic guided strategy.

merge feature vectors for meme understanding, and M3F (Wang et al. 2024) devise attention mechanisms to capture the interaction between text and images.

Main Results

We conduct a comprehensive comparison of our MGMCF with SoTA unimodal and multimodal models on the MET-MEME dataset. Table 1 presents the results for four tasks: sentiment analysis (SA), intention detection (ID), offensiveness detection (OD), and metaphor recognition (MR). The results highlight that our approach outperforms SoTA baselines across all evaluation metrics for the four tasks, revealing several key findings. Firstly, compared to unimodal models, multimodal models demonstrate superior performance by leveraging additional visual-textual features. However, it is crucial to thoroughly exploit visual information and deeply fuse multimodal features. Otherwise, they only yield marginal improvements over unimodal models or even underperform them (e.g., in the OD task, the MET method performs worse than DenseNet-161 and ResNet-50). By creating cross-modal attention, M3F achieves the current SoTA results. Notably, our model significantly surpasses the SoTA techniques. On the English meme dataset, our precision improves by 8.14% in the OD task, and accuracy improves by 3.53%, 3.89%, and 3.52% in MR, SA, and ID, respectively. On the Chinese dataset, our precision improves by 3.65%, 5.79%, 3.27%, and 5.6% in MR, SA, ID, and OD, respectively. Furthermore, our approach exhibits significant enhancements compared to MET. On the English MEME dataset, improvements range from 6.18% to 13.25%, and on the Chinese MEME dataset, enhancements range from 3.86% to 14.97%. These findings indicate that by focusing on object-level fine-grained image details and the intrinsic unimodal clues, and integrating multi-granular clues, we achieve significant performance improvements in the MMU.

Ablation Study

We perform ablation experiments to evaluate the contribution of each component in our model. As depicted in Table 2, no variant matches the full model's performance, highlighting the indispensability of each component. Specifically, when the global-local interaction is not utilized, three evaluation metric scores for all four tasks suffer the most significant decline. In particular, the precision score for the OD task drops by 5.16%, and for the SA task drops by 2.28%. This indicates that the global-local interaction successfully integrates information from different modalities, providing a more comprehensive multimodal feature representation. To validate the necessity of object-level semantic mining module, we remove this module, and the decline in results signifies its indispensable impact on MMU. This finding suggests that mining fine-grained information from images can offer more detailed insights into image content. Furthermore, removing unimodal predictions leads to a performance drop, indicating that unimodal predictions contribute to the final predictions and effectively address the issue of weak correlation between image and text. Additionally, performance declines when dual-semantic guided strategy is removed, demonstrating its crucial role in enhancing semantic alignment and reducing modalities' inconsistencies.

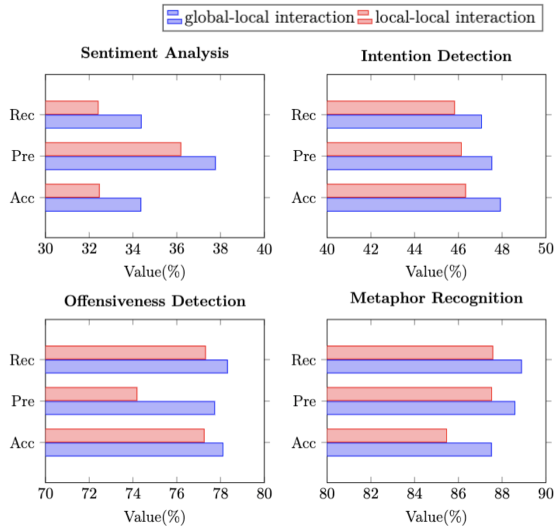


Figure 3: Comparative results between global-local and local-local interaction on the MET-MEME English dataset.

Deep Analyses on The Proposed Methods

To further investigate the effectiveness of our method, we conduct in-depth analyses to answer the following questions, aiming to mine the intuition and analyze implicit phenomena. **Q1: What are the advantages of the global-local interaction?** To further validate the effectiveness of our proposed global-local interaction model, we conduct a comparative analysis between our global-local interaction method and the local-local interaction method. The local-local interaction method refers to performing pairwise local interactions within each modality. As shown in Figure 3, the results consistently demonstrate the superiority of the global-local interaction across all evaluation metrics for the four tasks. This finding indicates that by incorporating a higher-level global context that encompasses the entire modality, the global-local interaction method achieves more comprehensive and effective interactions between modalities. In contrast, the local-

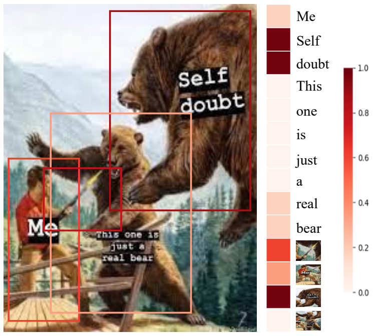


Figure 4: Visualization of a typical example.

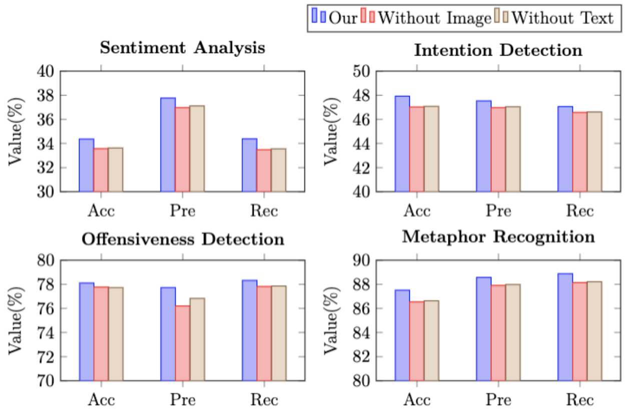


Figure 5: Influence of unimodal prediction on the MET-MEME English dataset.

local interaction method solely focuses on intra-modality local feature interactions and fails to fully leverage the holistic information across modalities. By leveraging the global context, we capture a broader range of semantic information, enabling a deeper understanding of the interdependence between modalities and effectively addressing inconsistencies and differences between modalities, thereby enhancing the performance of the MMU task.

Q2: How can fine-grained visual features help improve model performance? We conduct a visualization in Figure 4 to better illustrate the outstanding performance of the fine-grained visual enhancement module. By visualizing the attention distribution of the model, we observe that the module exhibits higher attention towards specific object parts in the image when the fine-grained visual enhancement module is utilized. Focusing on specific objects rather than the entire image allows for more accurate capture of crucial visual details. This confirms the effectiveness of the fine-grained visual enhancement module in improving the model's attention and understanding of key visual information.

Q3: What are the advantages of unimodal prediction in multimodal meme understanding? In order to validate the effectiveness of incorporating unimodal prediction in multimodal meme understanding, we conduct a comparative analysis of the performance of combining unimodal prediction with separately removing text modal prediction and image modal prediction. As illustrated in Figure 5, the results consistently show that the combined unimodal prediction outperform the scenarios where any unimodal prediction is removed across all evaluation metrics in the four tasks. This indicates that by focusing on the crucial information within each modality, we can achieve a more comprehensive and accurate understanding of the visual and textual elements within memes. Furthermore, we observe that removing the image modality prediction has a larger impact on the performance compared to removing the text modality prediction. This finding emphasizes the importance of enhancing visual information. By capturing fine-grained visual details, the model can better leverage the rich visual cues and context present in memes. These findings highlight the value of unimodal prediction and underscore the significance of considering the specific characteristics of each modality in MMU.

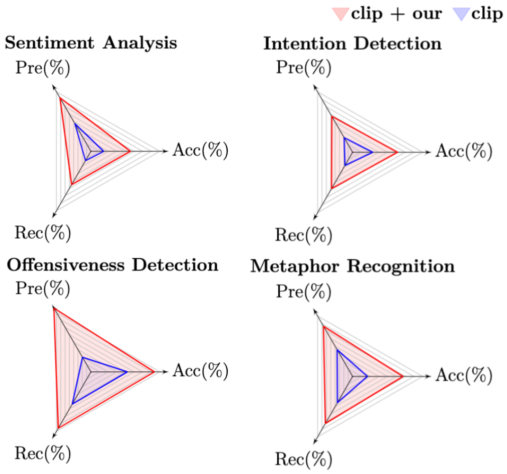


Figure 6: Influence of CLIP in MMU on the MET-MEME English dataset.

Q4: What impact does a large language model have on MMU? Considering the extensive usage of large language models (Wu et al. 2024a,b; Fei et al. 2024c,b), we investigate their influence on MMU. Figure 6 displays the results achieved by employing the standalone CLIP model and by integrating the CLIP model with our proposed method. Notably, employing the CLIP model alone yields impressive performance, attesting to its adeptness in comprehending multimodal memes. Moreover, the integration of the CLIP model with our method results in additional performance improvements, underscoring the efficacy of our approach.

Conclusion

In this paper, we explore two major challenges in the MMU task: the loss of fine-grained metaphorical visual clues and the neglect of weak correlation between multimodal text and images, proposing a solution named MGMCF. MGMCF enhances the complexity and diversity of images by capturing object-level fine-grained visual clues, and resolves the weak correlation between text and images through a novel global-local cross-modal interaction for multi-granular clue fusion. Extensive experiments on the MET-MEME bilingual dataset demonstrate the effectiveness of all our proposed innovative methods and hypotheses, achieving SoTA performance.

Acknowledgments

This work is supported by the National Key Research and Development Program of China (No. 2022YFB3103602), the National Natural Science Foundation of China (No. 62176187, No. 62202210). This work is also supported by the fund of Laboratory for Advanced Computing and Intelligence Engineering.

References

Alnajjar, K.; Hämäläinen, M.; and Zhang, S. 2022. Ring that bell: A corpus and method for multimodal metaphor detection in videos. *arXiv preprint arXiv:2301.01134*.

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6077–6086.

Badathala, N.; Kalarani, A. R.; Siledar, T.; and Bhattacharyya, P. 2023. A Match Made in Heaven: A Multi-task Framework for Hyperbole and Metaphor Detection. In Rogers, A.; Boyd-Graber, J. L.; and Okazaki, N., eds., *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, 388–401. Association for Computational Linguistics.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Elzohbi, M.; and Zhao, R. 2024. ContrastWSD: Enhancing Metaphor Detection with Word Sense Disambiguation Following the Metaphor Identification Procedure. In Calzolari, N.; Kan, M.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N., eds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, 3907–3915. ELRA and ICCL.

Fang, X.; Easwaran, A.; Genest, B.; and Suganthan, P. N. 2024a. Your Data Is Not Perfect: Towards Cross-Domain Out-of-Distribution Detection in Class-Imbalanced Data. *ESWA*.

Fang, X.; Fang, W.; Liu, D.; Qu, X.; Dong, J.; Zhou, P.; Li, R.; Xu, Z.; Chen, L.; Zheng, P.; et al. 2024b. Not all inputs are valid: Towards open-set video moment retrieval using language. In *Proceedings of ACM MM*.

Fang, X.; Liu, D.; Fang, W.; Zhou, P.; Xu, Z.; Xu, W.; Chen, J.; and Li, R. 2024c. Fewer Steps, Better Performance: Efficient Cross-Modal Clip Trimming for Video Moment Retrieval Using Language. In *Proceedings of AAAI*.

Fang, X.; Liu, D.; Zhou, P.; and Nan, G. 2023. You can ground earlier than see: An effective and efficient pipeline for temporal sentence grounding in compressed videos. In *Proceedings of CVPR*.

Fang, X.; Xiong, Z.; Fang, W.; Qu, X.; Chen, C.; Dong, J.; Tang, K.; Zhou, P.; Cheng, Y.; and Liu, D. 2025. Rethinking weakly-supervised video temporal grounding from a game perspective. In *Proceedings of ECCV*.

Fei, H.; Wu, S.; Ji, W.; Zhang, H.; Zhang, M.; Lee, M.-L.; and Hsu, W. 2024a. Video-of-thought: Step-by-step video reasoning from perception to cognition. In *Proceedings of the International Conference on Machine Learning*.

Fei, H.; Wu, S.; Zhang, H.; Chua, T.-S.; and Yan, S. 2024b. VITRON: A Unified Pixel-level Vision LLM for Understanding, Generating, Segmenting, Editing.

Fei, H.; Wu, S.; Zhang, M.; Zhang, M.; Chua, T.-S.; and Yan, S. 2024c. Enhancing video-language representations with structural spatio-temporal alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

He, X.; Yu, L.; Tian, S.; Yang, Q.; Long, J.; and Wang, B. 2024. VIEMF: Multimodal metaphor detection via visual information enhancement with multimodal fusion. *Inf. Process. Manag.*, 61(2): 103652.

Hee, M. S.; Chong, W.; and Lee, R. K. 2023. Decoding the Underlying Meaning of Multimodal Hateful Memes. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China*, 5995–6003. ijcai.org.

Heo, B.; Yun, S.; Han, D.; Chun, S.; Choe, J.; and Oh, S. J. 2021. Rethinking spatial dimensions of vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11936–11945.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708.

Ji, J.; Luo, Y.; Sun, X.; Chen, F.; Luo, G.; Wu, Y.; Gao, Y.; and Ji, R. 2021. Improving image captioning by leveraging intra-and inter-layer global representation in transformer network. In *Proceedings of the AAAI conference on artificial intelligence*, 1655–1663.

Ji, J.; Ma, Y.; Sun, X.; Zhou, Y.; Wu, Y.; and Ji, R. 2022. Knowing what to learn: a metric-oriented focal mechanism for image captioning. *IEEE Transactions on Image Processing*, 31: 4321–4335.

Ji, J.; Ren, W.; and Naseem, U. 2023. Identifying creative harmful memes via prompt based approach. In *Proceedings of the ACM Web Conference 2023*, 3868–3872.

Kiela, D.; Firooz, H.; Mohan, A.; Goswami, V.; Singh, A.; Ringshia, P.; and Testuggine, D. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33: 2611–2624.

Kirk, H. R.; Jun, Y.; Rauba, P.; Wachtel, G.; Li, R.; Bai, X.; Broestl, N.; Doff-Sotta, M.; Shtedritski, A.; and Asano, Y. M. 2021. Memes in the wild: Assessing the generalizability of the hateful memes challenge dataset. *arXiv preprint arXiv:2107.04313*.

Li, B.; Fei, H.; Liao, L.; Zhao, Y.; Teng, C.; Chua, T.-S.; Ji, D.; and Li, F. 2023a. Revisiting disentanglement and fusion on modality and context in conversational multimodal emotion recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, 5923–5934.

Li, J.; He, X.; Wei, L.; Qian, L.; Zhu, L.; Xie, L.; Zhuang, Y.; Tian, Q.; and Tang, S. 2022a. Fine-grained semantically aligned vision-language pre-training. *Advances in neural information processing systems*, 35: 7290–7303.

Li, J.; Xie, J.; Qian, L.; Zhu, L.; Tang, S.; Wu, F.; Yang, Y.; Zhuang, Y.; and Wang, X. E. 2022b. Compositional temporal grounding with structured variational cross-graph correspondence learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3032–3041.

Li, Y.; Wang, S.; Lin, C.; and Guerin, F. 2023b. Metaphor Detection via Explicit Basic Meanings Modelling. In Rogers, A.; Boyd-Graber, J. L.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, 91–100. Association for Computational Linguistics.

Lin, H.; Luo, Z.; Gao, W.; Ma, J.; Wang, B.; and Yang, R. 2024. Towards explainable harmful meme detection through multimodal debate between large language models. In *Proceedings of the ACM on Web Conference 2024*, 2359–2370.

Luo, M.; Fei, H.; Li, B.; Wu, S.; Liu, Q.; Poria, S.; Cambria, E.; Lee, M.-L.; and Hsu, W. 2024. Panosent: A panoptic sextuple extraction benchmark for multimodal conversational aspect-based sentiment analysis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 7667–7676.

Qiao, W.; Zhang, P.; and Ma, Z. 2024. A Quantum-Inspired Matching Network with Linguistic Theories for Metaphor Detection. In Calzolari, N.; Kan, M.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N., eds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, 1435–1445. ELRA and ICCL.

Qu, Y.; He, X.; Pierson, S.; Backes, M.; Zhang, Y.; and Zannettou, S. 2023. On the Evolution of (Hateful) Memes by Means of Multimodal Contrastive Learning. In *44th IEEE Symposium on Security and Privacy, SP 2023, San Francisco, CA, USA, May 21-25, 2023*, 293–310. IEEE.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Tan, M.; and Le, Q. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, 6105–6114. PMLR.

Tian, Y.; Xu, N.; Mao, W.; and Zeng, D. 2023. Modeling Conceptual Attribute Likeness and Domain Inconsistency for Metaphor Detection. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, 7736–7752. Association for Computational Linguistics.

Wang, B.; Huang, S.; Liang, B.; Tu, G.; Yang, M.; and Xu, R. 2024. What do they "meme"? A metaphor-aware multi-modal multi-task framework for fine-grained meme understanding. *Knowledge-Based Systems*, 294: 111778.

Wang, Z.; Mayhew, S.; Roth, D.; et al. 2019. Cross-lingual ability of multilingual bert: An empirical study. *arXiv preprint arXiv:1912.07840*.

Wu, S.; Fei, H.; Cao, Y.; Bing, L.; and Chua, T.-S. 2023. Information Screening whilst Exploiting! Multimodal Relation Extraction with Feature Denoising and Multimodal Topic Modeling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 14734–14751.

Wu, S.; Fei, H.; Li, X.; Ji, J.; Zhang, H.; Chua, T.-S.; and Yan, S. 2024a. Towards Semantic Equivalence of Tokenization in Multimodal LLM. *arXiv preprint arXiv:2406.05127*.

Wu, S.; Fei, H.; Qu, L.; Ji, W.; and Chua, T.-S. 2024b. NExT-GPT: Any-to-Any Multimodal LLM. In *Proceedings of the International Conference on Machine Learning*, 53366–53397.

Xu, B.; Li, T.; Zheng, J.; Naseriparsa, M.; Zhao, Z.; Lin, H.; and Xia, F. 2022. MET-Meme: A Multimodal Meme Dataset Rich in Metaphors. In Amigó, E.; Castells, P.; Gonzalo, J.; Carterette, B.; Culpepper, J. S.; and Kazai, G., eds., *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, 2887–2899. ACM.

Xu, N.; Zeng, Z.; and Mao, W. 2020. Reasoning with multimodal sarcastic tweets via modeling cross-modality contrast and semantic association. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, 3777–3786.

Yang, C.; Zhu, F.; Han, J.; and Hu, S. 2023. Invariant Meets Specific: A Scalable Harmful Memes Detection Framework. In *Proceedings of the 31st ACM International Conference on Multimedia*, 4788–4797.

Zhang, S.; and Liu, Y. 2023. Adversarial Multi-task Learning for End-to-end Metaphor Detection. In Rogers, A.; Boyd-Graber, J. L.; and Okazaki, N., eds., *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, 1483–1497. Association for Computational Linguistics.

Zheng, L.; Chen, B.; Fei, H.; Li, F.; Wu, S.; Liao, L.; Ji, D.; and Teng, C. 2024a. Self-Adaptive Fine-grained Multimodal Data Augmentation for Semi-supervised Muti-modal Coreference Resolution. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 8576–8585.

Zheng, L.; Fei, H.; Li, F.; Li, B.; Liao, L.; Ji, D.; and Teng, C. 2024b. Reverse multi-choice dialogue commonsense inference with graph-of-thought. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 19688–19696.

Zheng, L.; Ji, D.; Li, F.; Fei, H.; Wu, S.; Li, J.; Li, B.; and Teng, C. 2023a. ECQED: emotion-cause quadruple extraction in dialogs. *arXiv preprint arXiv:2306.03969*.

Zheng, L.; Li, F.; Chai, Y.; Teng, C.; and Ji, D. 2023b. A Bi-directional Multi-hop Inference Model for Joint Dialog Sentiment Classification and Act Recognition. In *CCF International Conference on Natural Language Processing and Chinese Computing*, 235–248. Springer.